ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

面向人民日报语料的新闻自动摘要生成

梁媛[1,2] 王东波[1,2] 黄水清[1,2]
1. 南京农业大学信息管理学院 南京 210095
2. 南京农业大学人文与社会计算研究中心 南京 210095

摘要：[目的/意义]面向主流新闻媒体人民日报语料展开研究，旨在为文本自动摘要研究提供思路和实践支撑，进而应用到新闻等相关文本信息处理中，为知识聚合服务和信息获取途径研究做出贡献。[方法/过程]以新时代人民日报语料 NEPD 中的 2015 年 1 月、2015 年 6 月和 2016 年 1 月的人民日报分词语料作为实验语料，基于 TF-IDF、Textrank 等抽取式自动摘要算法，以及基于指针生成网络的生成式自动摘要模型展开研究，并对摘要结果进行分析评价。[结果/结论]实验设计面向人民日报语料的新闻抽取式自动摘要算法，构建面向人民日报语料的新闻生成式自动摘要指针生成网络模型，并通过 Rouge 指标（包括 Rouge-1、Rouge-2 和 Rouge-L 3 种指标）对实验结果进行评测，为人民日报分词语料的应用提供具体思路，并对新闻自动摘要系统研究提供语料支持和实践支撑。

关键词：人民日报　抽取式自动摘要　生成式自动摘要　NEPD　指针生成网络

分类号：G255.1

引用格式：梁媛，王东波，黄水清. 面向人民日报语料的新闻自动摘要生成 [J/OL]. 知识管理论坛，2022, 7(4): 452-464[引用日期]. http://www.kmf.ac.cn/p/307/.

1 引言

网络信息的爆炸式增长在使人们获取信息更加便利的同时，也带来了信息利用效率低、阅读成本过高等问题，而自动摘要技术通过对信息的压缩和精炼，为提高知识获取效率提供了辅助手段[1]，该技术的产生和发展使解决上述问题成为可能。目前，自动摘要的主要方式有抽取式和生成式两种，抽取式自动摘要起步较早，经过许多学者多年研究，该技术已较为成熟，而随着机器学习引入到自动摘要领域，生成式自动摘要再一次迎来了发展的可能。

新闻是记录社会问题、传播时代信息、获取时事热点的重要途径，而《人民日报》是中国共产党中央委员会机关报，是国家与人民沟通的主要媒介，也是国内外文化交流的桥梁，因此，人民日报语料的研究具有重要意义。本文实验语料来自新时代人民日报语料库（New

作者简介：梁媛，博士研究生；王东波，教授，博士，博士生导师；黄水清，教授，博士，博士生导师，通信作者，E-mail：sqhuang@njau.edu.cn。

收稿日期：2022-04-25　　发表日期：2022-08-23　　本文责任编辑：刘远颖

知识管理论坛，2022（4）:452-464
DOI: 10.13266/j.issn.2095-5472.2022.038

Era People’s Daily Segmented Corpus，简称NEPD）[2]，NEPD 中收录的《人民日报》文章经过人工分词和校对，是具有良好可用性的精语料 [3]。通过 NEPD 中的语料可快速便捷地计算词语及其频次，进而进行后续的数据预处理，完成相应的文本处理任务。

笔者结合人们新闻浏览趋势的变化，针对大量新闻文本需要精炼的特征，面向人民日报语料，实现抽取式新闻自动摘要算法和生成式自动摘要模型的构建，并对摘要结果进行评价，进而提高新闻信息使用效率，节省用户阅读成本，为文本自动摘要技术及其评价方法提供思路。

2 相关研究

早期，莫燕 [4] 和王永成 [5] 介绍了自动文献摘要和自动提取知识的思想和算法。之后，王永成和许慧敏 [6]、王知津 [7] 分别提出并设计了 OA 中文文献自动摘要系统和基于句子选择的自动文本摘要系统，并对中文文献自动摘要的历史、发展和意义进行了概述。史磊和王永成 [8] 则对英文文献自动摘要系统进行了研究。

在前人研究的基础上，文本自动摘要研究得以快速发展，各类算法推陈出新。熊娇等 [9]、张筱丹和胡学钢 [10]、刘星含和霍华 [11]、纪文倩等 [12]、曾哲军 [13]、刘静和肖璐 [14] 分别采用图模型、向量空间模型、互信息、连续 LexRank 算法、依存句法分析图模型对文本进行自动摘要处理。王帅等 [15] 采用基于图模型和循环神经网络模型两阶段的长文本自动摘要方法，在大规模金融长文本数据上进行了摘要生成实验；吴云等 [16] 提高与标题相似的特征词的词频，进而计算词频矩阵和句子相似度，得到了词句协同的自动摘要提取算法；陈晨等 [17] 应用词句协同排序提出了基于图模型的自动摘要算法；丁建立等 [18] 采用多维度词嵌入模式，基于双编码器融入双通道语义对短文本进行自动摘要任务；冯读娟等 [19] 同样基于双编码器网络结构构建了 CGAtten-GRU 模型，并在大规模中文短文本摘要中取得良好的效果；廖涛等 [20] 参考图结构表示提出了事件网络表示文本中的事件关系，进而进行文本自动摘要；徐馨韬等 [21] 改进了 TextRank 算法，将 Doc2Vec 模型和 K-means 算法融入其中，优化了主题句提取生成摘要的效果；陈海华等 [22] 将引文上下文内容特征与支持向量机（support vector machine, SVM）模型融合，对学术文本进行自动摘要；黄水清等 [23] 根据计算机类文献设计了该领域自动文本摘要系统；张晗和赵玉虹 [24] 则针对医学文本，对文本及语义关系进行规范化抽取和语义图的构建，以实现句子主题归类，进而生成摘要；陈志敏等 [25]、李芳和何婷婷 [26] 则从信息检索方面入手，基于用户查询扩展及查询文档集合辅助生成摘要。

在这些算法中，采用主题划分、多特征融合算法的自动摘要研究尤为突出。张哲铭等 [27] 提出了结合主题感知与通信代理的高质量长文本摘要模型，能够生成主题突出的摘要结果；陈燕敏等 [28] 提出了一种融合主题与内容的自动摘要方法，并通过指代消解获得具有良好的连贯性和流畅性的自动摘要结果；罗芳等 [29] 改进了图模型方法，基于隐含狄利克雷分布（latent Dirichlet allocation，LDA）主题模型挖掘出的主题语义信息，将主题特征、统计特征和句间相似度等多维度对文本进行度量和抽取，最终达到深层主题语义挖掘利用的目的，实现自动摘要；杜秀英 [30] 针对大规模多文本摘要，构建了基于聚类与语义相似分析的 MapReduce 自动摘要架构，在时间性能、压缩效果和摘要质量上都有一定的提升。但以上方法和模型主要集中于抽取式自动摘要的研究，而对于生成式自动摘要仍有较大的研究空间。

随着大数据和人工智能技术的迅猛发展，传统自动文摘研究正朝着从抽取式摘要到生成式摘要的方向演化，从而达到生成更高质量的自然流畅的文摘的目的。近年来，深度学习技术逐渐被应用于生成式摘要研究中。吴世鑫等 [31] 基于带注意力、Pointer 机制和 Coverage 机制的 Sequence-to-Sequence 模型引入语义对齐的神经

网络，实现生成式自动摘要模型的构建；方旭等[32]提出了一种结合核心词修正的长短期记忆网络（long short-term memory，LSTM）算法自动生成中文短文本摘要；唐晓波和翟夏普[33]改进了 PageRank 算法，并采用句子向量化、分类器分类、句群划分和句子重组混合机器学习模型进行多文档自动摘要研究；谭金源等[34]和张克君等[35]融合多个深度学习模型分别提出了 Bi-MulRnn+ 和 BERT- 指针生成网络 BERT-PGN 生成式自动摘要模型，有效改善了生成式摘要的准确性和流畅度；李维勇等[36]、肖元君和吴国文[37]也都进行了基于深度学习的中文生成式自动摘要模型的研究与实现。

逐渐加快的生活节奏不断改变着人们的阅读习惯，人们从纸质书籍、报刊转向电子化阅读，阅读的新闻也逐渐转为短文本，因此，新闻媒体以及读者对于新闻摘要自动化的需求也随之增大。官礼和[38]分析了中文网络新闻自动摘要的思路和流程，并通过实验进行了分析佐证；韩永峰等[39]探讨了自动摘要中信息冗余的问题，并提出了基于事件抽取的网络新闻多文档自动摘要的改进方法；沈洲等[40]建立了新闻文献主题提取规则库，构建了面向新闻文献基于规则的自动摘要系统；李孟爽等[41]提出的自动摘要算法是基于互信息对文本词句语义特征的计算结果，并据此进行主题划分，抽取出关键句生成最终的文本摘要；王凯祥和任明[42]为满足用户查询的信息需求，设计了基于查询的新闻自动摘要算法，还与 TF-IDF、TextRank、LDA 等 6 种方法进行了对比实验；黄小江等[43]基于协同图排序模型自动生成了新闻话题的对比摘要，具有很强的新颖性；柯修和王惠临[44]则融合多种算法，包括指代消解、文本外部特征和图排序方法，实现了汉语、英语、孟加拉语 3 个语种的多文档新闻自动摘要；叶雷等[45]同样采用图排序方法，提出了多特征融合的汉越双语新闻摘要方法，能够自动获取同一事件的汉越双语新闻摘要。除新闻外，如微博、论坛等用户自主生成内容中的信息也拥有巨大的研究价值，而自动摘要是获取这类重要信息的一种手段，但这些短文本高冗余、高噪声等特征对于自动摘要造成较大的影响[46]，学者们[47-50]也在为解决这一问题作出不懈努力。

通过对上述文献的梳理可以发现，从基于规则、基于统计到后来的深度学习，从普通文本到动态视频，自动摘要技术的研究正随着技术的进步和用户的需求不断更迭发展着。而新闻自动摘要一直具有重要意义，其能够在很大程度上满足人们快节奏生活中的新闻获取。但目前新闻自动摘要的应用型研究主要集中在新闻的抽取式自动摘要上，而对于生成式自动摘要尚未有领域性、准确性较强的模型和系统。因此，笔者面向人民日报语料展开自动摘要的研究，通过传统算法和深度学习算法完成自动摘要任务，旨在根据当前主流新闻媒体的文本特征构建自动摘要模型，解决用户阅读长文本新闻耗时长、信息利用率低的问题，同时也为新闻媒体的知识聚合服务提供帮助，为新闻传播、文化传承提供新思路。

3 算法模型介绍

自然语言处理（natural language processing，NLP）作为一个传统研究领域，自其产生始终热度不减，其中缘由不只是新技术的诞生和引入，也因 NLP 有“最困难的人工智能子领域”之名。其中的自动摘要任务也是研究者们不断研究、突破的主要难点之一，特别是在快速阅读成为人们生活中非常重要的阅读方式的前提下。目前，自动摘要方法按生成方式主要分为抽取式自动摘要和生成式自动摘要，抽取式自动摘要主要应用关键词句排序的思想，而生成式自动摘要更多是基于深度学习模型来完成。在本文的实验中，抽取式自动摘要主要运用了关键词确定句子权重和 TextRank 等传统算法的思想，生成式自动摘要则参考了基于指针生成网络构建的面向中文的 Text-Summarizer-Pytorh-Chinese 模型[50]及其思路。

ChinaXiv合作期刊
知识管理论坛，2022（4）:452-464
DOI: 10.13266/j.issn.2095-5472.2022.038

3.1 抽取式自动摘要

本研究中的抽取式自动摘要主要采用的是按词频和簇确定关键词，再通过关键词对所在句打分，分数排序确定最终生成摘要的句子。这种方法源自 IBM 公司 H. P. Luhn 的一篇文章 *The Automatic Creation of Literature Abstracts*[51]，他提出用簇（cluster）表示关键词的聚类结果，这里的簇即包含多个关键词的句子片段，如图 1 所示：



图 1 关键词簇聚类示意图

簇权重的计算公式[52]如下：

$$簇权重=\frac{簇中关键词数量^2}{簇长} \quad 公式（1）$$

其中，簇长指句子片段中所包含词语的数量，以本研究中的部分人民日报语料为例：

“经过全国各族人民共同努力，‘十二五’规划圆满收官，广大人民群众有了更多获得感”，分词后语料实例为：

“经过 / 全国 / 各族 / 人民 / 共同 / 努力 /，/ ‘/ 十二五 /’ / 规划 / 圆满 / 收官 /，/ 广大 / 人民 / 群众 / 有 / 了 / 更 / 多 / 获得感”，

设“‘十二五’规划圆满收官”为一簇，簇长为 6，“十二五”“规划”“收官”为关键词，“广大人民群众有了更多获得感”为另一簇，簇长为 8，关键词为“人民”“群众”“获得感”，则两簇权重分别为 $3^2/6=1.5$ 和 $3^2/8=1.125$。按权重对文本包含的句子进行排序，确定抽取阈值（本文设定的阈值为 10，即抽出重要性最高的前 10 个句子），将这 10 个句子整合，即为该文本的自动摘要。类似 TextRank 算法，该算法源于 PageRank 算法，相当于将网页替换为句子，通过句子相似度矩阵以及设定的阈值来获得得分较高的句子作为自动摘要结果，这是一种无监督的抽取式自动摘要。

3.2 生成式自动摘要

指针生成网络（pointer-generator network）的自动摘要任务原理见图 2。该模型能够通过自注意力机制集中于文本中的重要词汇，并由此生成新词汇。同时，它不是通过复制原词来生成摘要，而是权衡词表中词汇的概率、词汇分布以及注意力分布来确定候选词的权重并获得最终分布情况。

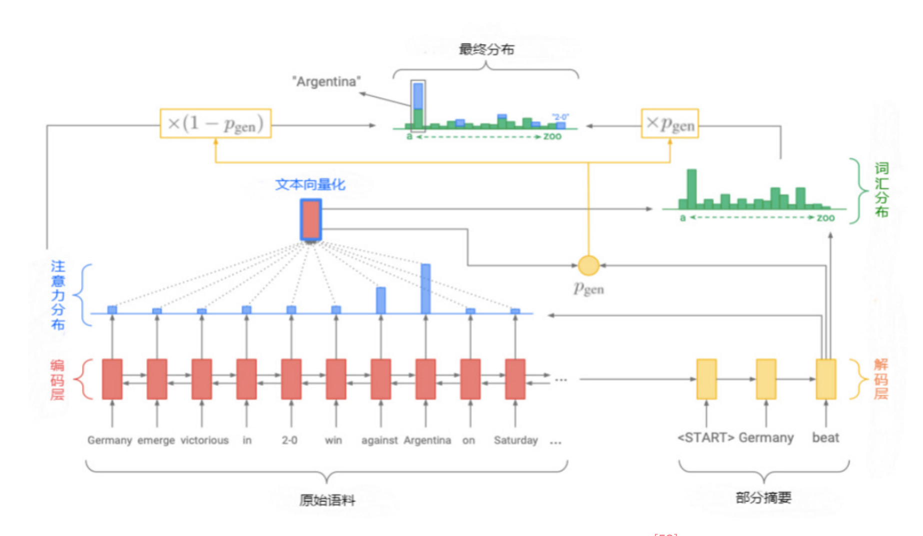


图 2 指针生成网络自动摘要原理图示[53]



知识管理论坛
2022 年 第 4 期(总第 40 期)

目前，面向中文的基于指针生成网络自动摘要的模型较少，因此，笔者参考 Text-Summarizer-Pytorch-Chinese 的构建思路，将预训练语料调整为 NEPD 语料，词表也针对 NEPD 语料进行了更新，之后再进行预训练和模型构建。

4 面向人民日报语料的新闻自动摘要生成实验

“《人民日报》是一张权威、严肃的综合性日报，凭借其采编力量对新闻事件做出反应，报道国内外重大事件”[54]。作为耳目与喉舌、桥梁和纽带的主流媒体，其文本信息价值不言而喻，人民日报语料一直以来也是研究者们的重要数据来源，其中，北京大学计算语言学研究所构建的人民日报语料库[55]是我国第一个大型的现代汉语标注语料库，之后，南京农业大学人文与社会计算研究中心在 2019 年对 2015 年至 2018 年《人民日报》发表的文章进行加工处理，构建了新时代人民日报语料库（NEPD）[56]。本研究以 NEPD 中 2015 年 1 月、2015 年 6 月和 2016 年 1 月 3 个月的语料为实验对象展开研究，原始语料如图 3 所示：



201501.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

全国/政协/举行/新年/茶话会/
习/近平/发表/重要/讲话/李/克强/张/德江/刘/云山/王/岐山/张/高丽/出席/俞/正声/主持/
十二月/三/十/一/日/，/全国/政协/在/北京/举行/新年/茶话会/。/
党/和/国家/领导人/习/近平/、/李/克强/、/张/德江/、/俞/正声/、/刘/云山/、/王/岐山/、/张/高丽/出席/茶话会/并/观看/演出/。/
新华社/记者/（/李/学仁/）/摄/
新华社/北京/12/月/31/日/电/（/记者/吴/晶晶/、/孙/铁翔/）/中国/人民/政治/协商/会议/全国/委员会/12/月/31/日/上午/在/全国/政协/礼堂/举行/新年/茶话会/。/
党/和/国家/领导人/习/近平/、/李/克强/、/张/德江/、/俞/正声/、/刘/云山/、/王/岐山/、/张/高丽/等同/各/民主党派/中央/、/全国/工商联/负责人/和/无党派/人士/代表/、/中央/和/国家/机关/有关/方面/负责人/以及/首都/各族/各界/人士/代表/欢聚一堂/，/共/迎/2015/年/元旦/。/
中共/中央/总书记/、/国家/主席/、/中央/军委/主席/习/近平/在/茶话会/上/发表/重要/讲话/。/
他/强调/，/问题/是/时代/的/声音/，/人心/是/最大/的/政治/。/
推进/党/和/国家/各项/工作/，/必须/坚持/问题/导向/，/倾听/人民/呼声/。/
我们/的/目标/越/伟大/，/我们/的/使命/越/艰巨/，/就/越/需要/所有/人/拧/成/一/股/绳/去/干事/创业/。/
让/我们/更加/紧密/地/团结/起来/，/向/着/我们/共同/的/奋斗/目标/、/向/着/更加/辉煌/的/明天/奋勇前进/。/
习/近平/代表/中共/中央/、/国务院/和/中央军委/，/向/各/民主党派/、/工商联/和/无党派/人士/、/各/人民团体/，/向/全国/广大/工人/、/农民/、/知识分子/、/干部/和/各界/人士/，/向/人民/解放军/指战员/、/武警/官兵/和/公安/干警/，/向/香港/特别/行政区/同胞/、/澳门/特别/行政区/同胞/、/台湾/同胞/和/海外/侨胞/，/向/关心/和/支持/中国/现代化/建设/的/国际/友人/，/致以/节日/的/祝福/，/祝/大家/新年/好/。/
习/近平/指出/，/在/过去/的/一/年/里/，/中共/中央/团结/带领/全国/各族/人民/，/坚持/稳中求进/工作/总/基调/，/积极/适应/经济/发展/新/常态/，/注重/谋划/全局性/、/战略性/、/长远性/的/重大/问题/，/推动/社会主义/经济/建设/、/政治/建设/、/文化/建设/、/社会/建设/、/生态/文明/建设/以及/国防/和/军队/建设/、/外交/工作/取得/重大/进展/。/
中共/中央/重点/抓/了/党/的/群众/路线/教育/实践/活动/，/聚焦/惩治/形式/主义/、/官僚/主义/、/享乐/主义/和/奢靡/之/风/，/党风/政风/为之一新/。/
我们/加大/反/腐败/斗争/力度/，/坚持/“/老虎/”/、/“/苍蝇/”/一起/打/，/一/批/腐败/分子/被/绳/之/以/党纪国法/。/
一/年/来/，/我们/蹄/疾/步/稳/地/推进/各项/改革/，/中央/全面/深化/改革/领导/小组/确定/的/80/个/重点/改革/任务/基本/完成/，/此外/中央/有关/部门/还/完成/了/108/个/改革/任务/，/各/方面/共/出台/370/条/改革/成果/，/一些/多年/来/难/啃/的/硬骨头/啃/下来/了/，/改革/为/我国/发展/注入/了/强大/动力/。/
这些/成绩/，/是/大家/共同/创造/的/，/光荣/属于/大家/。/
习/近平/强调/，/当前/，/时/和/势/总体/有利/，/但/艰和险/在/增多/。/
我们/要/全面/贯彻/落实/中共/十八大/和/十八/届/三中/、/四中/全会/精神/，/以/邓小平理论/、/“三个代表”重要思想/、/科学发展观/为/指导/，/继续/推进/全面/建成/小康/社会/、/全面/深化/改革/、/全面/依法/治国/、/全面/从严/治党/，/突出/创新/驱动/，/强化/风险/防控/，/加强/民生/保障/，/如期/完成/“/十二五/”/规划/确定/的/各项/目标/

图 3 NEPD 原始语料截图示例

4.1 数据预处理

根据本研究需要，笔者将每篇新闻从源语料中分割出来，处理后的文本见图 4，为之后的摘要抽取和生成做准备。经过数据清洗（同时清洗了未生成标准摘要的数据），获得 2015 年 1 月新闻 2 628 条、2015 年 6 月新闻 916 条、2016 年 1 月新闻 2 748 条，共计 6 292 条数据，本研究将以上述数据作为研究对象进行自动摘要研究。

4.2 实验环境与参数设置

本实验中生成式自动摘要模型训练及测试时采用的操作系统为 ubuntu 16.04，内存为 16GB DDR4，显存为 4GB GDDR5，CPU 为 Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz，GPU 型号为 NVIDIA Quadro K1200。生成式自动摘要模型参数设置如表 1 所示。

ChinaXiv合作期刊
知识管理论坛，2022（4）:452-464
DOI: 10.13266/j.issn.2095-5472.2022.038

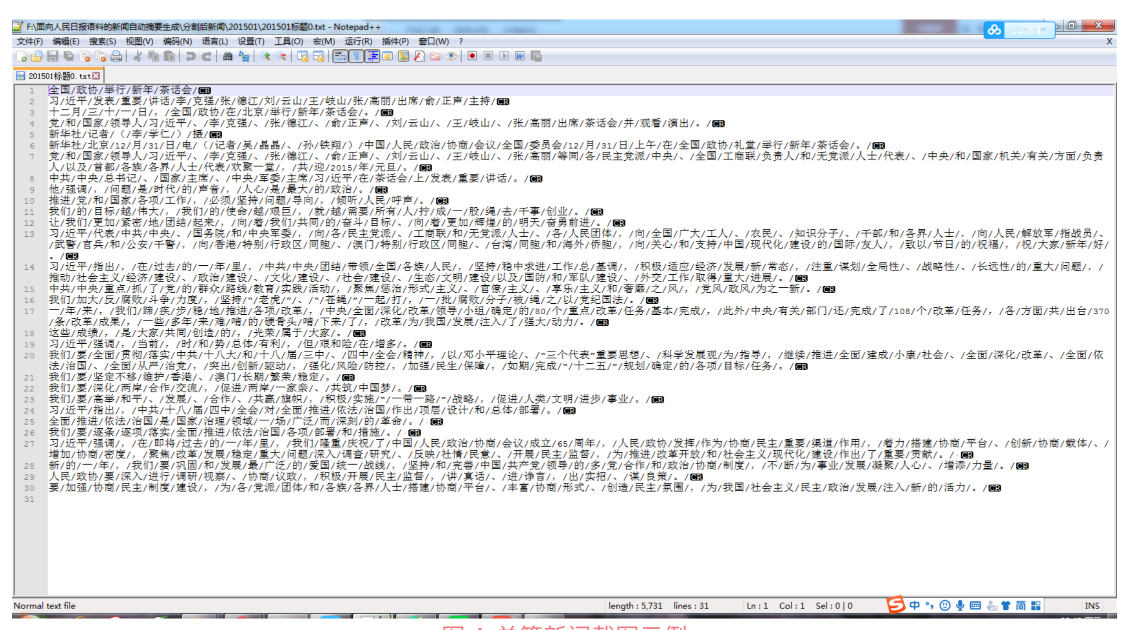


图 4 单篇新闻截图示例

表 1 生成式自动摘要模型参数设置

模型	参数设置
生成式自动摘要模型	hidden_dim = 512 emb_dim = 256 batch_size = 200 max_enc_steps = 100 max_dec_steps = 20 beam_size = 4 min_dec_steps = 3 vocab_size = 40000 rand_unif_init_mag = 0.02 trunc_norm_init_std = 1e-4 eps = 1e-12 max_iterations = 5000000

4.3 实验流程

本研究主要分为两个部分：面向人民日报语料的新闻抽取式自动摘要算法（以下简称“抽取式自动摘要算法”）研究，以及面向人民日报语料的新闻生成式自动摘要模型（以下简称“生成式自动摘要模型”）构建。

在抽取式自动摘要算法实验中，主要包括以下 8 个步骤：①人民日报分词语料获取；②待摘要文本预处理：包括去除特殊字符和空格空行等；③去停用词和词频统计：由于本研究选用的 NEPD 语料为精校过的分词语料，因此，不需要进行分词处理，在去停用词后直接进行词频统计即可；④计算句子权重：参考特征包括标题关键词信息、句子长度等特征；⑤根据权重对句子进行排序；⑥选定合适的阈值提取摘要句；⑦生成摘要；⑧根据标准摘要对自动摘要进行评价（评价指标包括 Rouge-1、Rouge-2 和 Rouge-L）。

生成式自动摘要模型构建过程主要包括以下 7 个步骤：①人民日报分词语料获取；②待摘要文本预处理：去除特殊字符和空格空行等，并根据模型要求调整训练语料格式；③预训练模型构建：将步骤②中的语料进行预训练，得到具有《人民日报》特色的预训练模型；④加入特征：根据 NEPD 分词语料统计关键词，并作为自定义词表引入到模型训练中，同时加入标题特征；⑤生成式自动摘要模型训练：根据训练过程及结果调整参数并进行迭代训练；⑥根据最终模型生成摘要；⑦根据标准摘要对自动摘要进行评价（评价指标包括 Rouge-1、Rouge-2 和 Rouge-L）。

5 实验结果评价与分析

由于目前尚无针对人民日报语料的摘要标

准语料库，因此，笔者在对自动摘要实验结果进行评价时，分别以关键词词频抽取式自动摘要结果和百度智能云的新闻摘要接口的分析结果作为标准摘要集合。百度智能云的新闻摘要是基于深度语义分析模型自动抽取文本，能够根据文本中的关键信息进一步生成指定长度的新闻摘要 [57]。

以本文选取的人民日报语料为例：

标准摘要（关键词词频抽取式自动摘要）：

“恐怖主义是国际社会公敌，中国历来反对一切形式的恐怖主义，积极参与国际反恐合作。军队和武警部队出境执行反恐任务，要遵守《联合国宪章》的宗旨和原则，遵循国际关系准则，并充分尊重当事国的主权。至于今后军队和武警部队是否赴境外反恐，将根据国家统一部署作出安排。”

自动摘要（面向人民日报语料的抽取式自动摘要）：

“中国军队和武警部队赴境外反恐将根据国家统一部署作出安排，军队和武警部队出境执行反恐任务，要遵守《联合国宪章》的宗旨和原则，遵循国际关系准则，并充分尊重当事国的主权。至于今后军队和武警部队是否赴境外反恐，将根据国家统一部署作出安排。”

标准摘要（百度智能云新闻摘要）：

“据报道，在河南省南阳市镇平县城郊乡的大刘营村，因当地污染严重，怀孕的村民只能离村待产。媒体曝光之后，当地已经责令涉事企业停产整治，并且问责环保部门领导。村民以这种方式远离环境污染，映射出对美好生态环境的要求底线，更映射出恶意排污的现实和环保执法的缺位。让我们的后代成长在美好的环境中，这是我们对子孙后代的责任。”

自动摘要（面向人民日报语料的生成式自动摘要）：

“重庆的不会愿意折腾到外村村民村民以这种方式远离环境污染映射出对美好的生态环境。”

5.1 评价指标

Rouge(recall-oriented understudy for gisting evaluation) 是评估自动摘要、机器翻译等自然语言处理任务的常用指标，它是将标准摘要和自动生成摘要进行相似度计算，得到的数值即为评价结果，计算公式如下 [58]：

$$Rouge-N=\frac{\sum_{S\in\{\mathrm{Re\,feremceSummaries}\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{\mathrm{Re\,feremceSummaries}\}}\sum_{gram_n\in S}Count(gram_n)}$$

公式（2）

其中，分母为 n-gram 个数，分子为标准摘要和自动摘要共有的 n-gram 个数。例如，Rouge-1 中的分子是自动摘要和标准摘要中均出现的 1-gram 的个数，分子是标准摘要的 1-gram 个数。笔者选取的评价指标为 Rouge-1、Rouge-2 和 Rouge-L，Rouge-L 是指运用 LCS（longest common subsequence，最长公共子序列）计算的 Rouge 评测指标，计算公式分别为：

$$R_{lcs}=\frac{LCS(X,Y)}{m}$$ 公式（3）

$$P_{lcs}=\frac{LCS(X,Y)}{n}$$ 公式（4）

$$F_{lcs}=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}}$$ 公式（5）

其中，$LCS(X,Y)$ 是 X 和 Y 的最长公共子序列的长度，m 和 n 分别表示标准摘要和自动摘要的长度（通常为词语个数），R_{lcs} 和 P_{lcs} 分别表示召回率和准确率。Β 的数值通常较大，导致 Rouge-L 几乎只考虑召回率 R_{lcs}，这与 Rouge-N 相同。

此外，上述 3 种 Rouge 评测指标内部运用的 P、R、F 为准确率(Precision)、召回率(Recall)、F 值（F-Measure）。具体计算公式分贝如下：

$$准确率P=\frac{正确识别的句对}{正确识别的句对+被错误识别的句对}\times100\%$$ 公式 (6)

$$召回率R=\frac{正确识别的句对}{正确识别的句对+未被识别的句对}\times100\%$$ 公式 (7)

$$调和平均值F=\frac{2\times P\times R}{P+R}\times100\%$$ 公式 (8)

ChinaXiv合作期刊
知识管理论坛，2022（4）:452-464
DOI: 10.13266/j.issn.2095-5472.2022.038

5.2 实验结果

在抽取式自动摘要实验中，本研究分别通过词频和簇聚类抽取关键词的方式对句子进行打分，并按分数对句子进行排序，进而抽取出相应的摘要结果。将词频抽取式自动摘要结果作为标准摘要，将簇聚类抽取式自动摘要作为自动摘要结果并与标准摘要进行 Rouge 评测，部分摘要结果截图如图 5 所示：

1	p	r	f	p	r	f	p	r	f
2	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
3	0.541666667	1	0.702702698	0.491525424	1	0.659090905	0.541666667	1	0.702702698
4	0.950819672	1	0.974789911	0.917808219	1	0.957142852	0.950819672	1	0.974789911
5	1	0.57337884	0.728850321	1	0.541436464	0.702508956	1	0.57337884	0.728850321
6	0.813953488	1	0.897435892	0.765822785	1	0.867383508	0.813953488	1	0.897435892
7	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
8	0.864197531	1	0.927152313	0.81512605	1	0.898148143	0.864197531	1	0.927152313
9	0.931623932	1	0.964601765	0.871794872	1	0.931506844	0.931623932	1	0.964601765
10	1	0.765957447	0.867469875	1	0.719745223	0.837037032	1	0.765957447	0.867469875
11	0.722689076	1	0.839024385	0.685185185	0.991071429	0.810218973	0.722689076	1	0.839024385
12	1	0.887755102	0.940540536	1	0.85620915	0.922535206	1	0.887755102	0.940540536
13	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
14	0.734693878	1	0.847058819	0.694656489	1	0.819819815	0.734693878	1	0.847058819
15	1	0.775	0.873239432	1	0.708661417	0.829493083	1	0.775	0.873239432
16	1	0.80620155	0.892703858	1	0.775956284	0.873846149	1	0.80620155	0.892703858
17	0.6625	1	0.796992476	0.594339623	1	0.745562126	0.6625	1	0.796992476
18	1	0.463414634	0.633333329	1	0.374732334	0.545171336	1	0.463414634	0.633333329
19	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
20	0.708860759	1	0.829629625	0.605504587	1	0.75428571	0.708860759	1	0.829629625
21	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
22	1	0.634146341	0.776119398	1	0.607526882	0.755852838	1	0.634146341	0.776119398
23	0.413461538	1	0.585034009	0.346405229	1	0.514563103	0.413461538	1	0.585034009
24	1	0.781954887	0.877637126	1	0.74611399	0.854599402	1	0.781954887	0.877637126
25	0.568	1	0.724489791	0.490291262	1	0.657980452	0.568	1	0.724489791
26	0.884210526	1	0.938547481	0.84057971	1	0.913385822	0.884210526	1	0.938547481
27	1	1	0.999999995	1	1	0.999999995	1	1	0.999999995
28	1	0.519230769	0.683544299	1	0.477941176	0.646766165	1	0.519230769	0.683544299
29	0.714285714	1	0.833333328	0.646666667	1	0.785425096	0.714285714	1	0.833333328
30	1	0.356020942	0.525096521	0.994736842	0.278350515	0.434982735	1	0.356020942	0.525096521

图 5 自动摘要实验结果示例

全部自动摘要的综合评测结果见表 2。通过表 2 可以看出，整体上抽取式自动摘要实验结果抽取效果良好（均值：Rouge-1=0.8447，Rouge-2=0.8257，Rouge-L=0.8446），能够对原始语料进行大致概括。由于在抽取式自动摘要实验中，标准摘要同样为自动生成，且在 Rouge 指标计算相似度的过程中，一旦抽取出的语句与标准摘要不同，则两个对应的完整长句相似度将会极低，这可能会导致 Rouge 指标明显偏低的问题出现。因此，笔者将会在未来的研究中一方面调整标准摘要的准确度，另一方面完善自动摘要的评价方法。

表 2 抽取式自动摘要实验评测结果

评价指标	Rouge-1			Rouge-2			Rouge-L		
	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%
201501	88.55	88.69	85.17	86.59	87.13	82.33	88.54	88.69	85.16
201506	90.48	75.66	76.23	88.85	73.05	72.32	90.44	75.64	76.20
201601	87.69	89.06	84.82	85.65	87.52	81.92	87.69	89.05	84.81
均值	88.91	84.47	82.07	87.03	82.57	78.86	88.89	84.46	82.06

在生成式自动摘要实验中，本研究将全部新闻语料进行预处理，接入百度智能云新闻摘要接口，获取相应的自动摘要结果，由于该平台输入文本长度有限，因此，笔者经过代码筛选，共获得 7 967 条符合文本长度限制的新闻文本。另外，由于本文拟构建的指针生成网络模型需要大规模训练语料，遂将 2015 年 1 月、2015 年 6 月和 2016 年 1 月 3 个月的原始语料合并后再继续进行实验。

对语料进行预处理后，将原始文本和标准摘要（百度智能云生成摘要）匹配并输入指针生成网络模型中进行训练和测试。笔者在模型训练过程中引入了自定义词表，该词表由 NEPD 分词语料生成，能够提高自动摘要模型的训练效果，以及生成摘要的流畅度和贴合度。在结果评价阶段，笔者在生成式自动摘要实验中采用的同样是 Rouge 指标进行评价，评测结果如表 3 所示：

表 3 生成式自动摘要实验评测结果

模型	Rouge-1/%	Rouge-2/%	Rouge-L/%
0005000	26.06	14.99	24.46
0010000	26.10	15.05	24.47
0015000	26.05	15.46	24.51
0020000	22.47	12.15	20.91
0025000	3.82	0.10	3.68
0030000	3.83	0.09	3.67
0035000	3.82	0.09	3.67
0040000	3.82	0.06	3.67

以本研究生成结果的其中一组数据为例(见表 4），不同算法或模型生成的摘要内容有一定的差别，但总体上流畅度问题较小，可读性有一定的差别。抽取式自动摘要由于单句抽取自人民日报原文，因此句子内部可读性高于生成式摘要，句间连贯性低于生成式摘要。从摘要内容整体上看，抽取式摘要包含的内容更丰富，但概括能力较差，内容冗余，句子间关联度较低；而生成式自动摘要有一定的语义理解能力，生成的摘要内容更简练，相对比较符合新闻摘要的特征，对原始语料的总结更灵活，但会出现个别词汇重复、摘要内容不全面等问题。

本研究选用的评测指标为 Rouge 指标，这种评价方式虽然直观简洁并且能够在一定程度上反映词序和摘要效果，但该指标区分度不高，特别是 Rouge-N 中 N>3 时，指标数值通常较小 [58]，对结果评价有较大影响。除此之外，Rouge 指标主要是根据文本相似度对标准摘要和自动摘要进行对比，同时，它具有一定的奖励机制，会给予原始表达（词汇）更高的分数 [53]，这就导致在同一篇新闻中，通常抽取式自动摘要的分数会高于生成式自动摘要。因此，这种计算方式有一定的局限性，特别是对于生成式自动摘要而言更是如此。笔者将在后续研究中尝试多种评价方式对实验结果进行综合测评，主要包括人工生成摘要数据，将其作为标准摘要数据集，或通过对生成的自动摘要人工打分的方式进行评价，以求得到更准确的评价结果。

ChinaXiv合作期刊
知识管理论坛，2022（4）:452-464
DOI: 10.13266/j.issn.2095-5472.2022.038

表 4 面向《人民日报》的新闻自动摘要生成结果样例

数据样例	文本内容
原始语料	世界经济论坛将日益成为连接中国与世界的桥梁，成为增进世界对中国发展道路了解、认同与支持的重要平台每年1月，全球目光聚焦瑞士东部小镇达沃斯——来自世界各地的政、商、学界人士群贤毕至、济济一堂，共商天下大事。即将于1月21日至24日召开的第四十五届世界经济论坛2015年年会将是盛况空前，近50位国家元首和政府首脑将出席，来自140多个国家的2500多位各界精英，将就世界面临的政治、经济、社会、科技等问题探讨应对之道。今年达沃斯年会的主题是“全球新局势”。当今世界正经历深刻复杂变化，国际体系和国际秩序深度调整，国际力量对比深刻变化，各种新机遇新挑战层出不穷。用世界经济论坛创立者、执行主席施瓦布教授的话说，“世界正处在一个十字路口，2015年将是决定未来命运的关键之年”。值此之际，人们都想知道，中国对当前的“全球新局势”有何真知灼见，有何治理良方，中国会采取什么样的应对之道。人们还想知道，作为世界第二大经济体的中国，经济发展前景如何，在推动国际经济合作上将发挥什么样的作用。年会上的中国声音，必将增强国际社会对中国发展的信心，也会增强对世界经济前景的信心。中国通过全面深化改革，推动经济社会持续健康发展，将为世界提供更多机遇。中国推动建立以合作共赢为核心的新型国际关系，必将使全世界从中受益。中国的发展需要一个和平稳定的世界，中国的发展也必定促进世界的和平与稳定。中国与世界经济论坛的渊源可谓久远，早在改革开放之初，中国就与论坛建立了合作关系。上世纪90年代初以来，中国的多位国家领导人曾出席达沃斯年会。每年9月世界经济论坛在中国召开的新领军者年会，又称夏季达沃斯年会，与冬季达沃斯年会并蒂相映、相得益彰。我本人有幸多次参加冬季和夏季达沃斯年会，还参与过夏季达沃斯年会的筹组工作。我与施瓦布教授相识多年，结下了深厚的友谊。去年初，我抵日内瓦履新，他是邀我餐叙的第一个外国人。施瓦布教授曾多次邀我商讨如何筹备今年的论坛年会，还与我探讨如何进一步深化论坛与中方的合作。当前中方与世界经济论坛合作势头良好，相信将来扩大合作的前景十分广阔。世界经济论坛将日益成为连接中国与世界的桥梁，成为增进世界对中国发展道路了解、认同与支持的重要平台。当下，达沃斯小镇银装素裹，国际媒体已开始将镜头聚焦小镇，世界对即将在达沃斯发出的中国声音充满期待。（作者为中国常驻联合国日内瓦办事处和瑞士其他国际组织代表吴海龙）
抽取式 标准摘要	人们还想知道，作为世界第二大经济体的中国，经济发展前景如何，在推动国际经济合作上将发挥什么样的作用。中国与世界经济论坛的渊源可谓久远，早在改革开放之初，中国就与论坛建立了合作关系。每年9月世界经济论坛在中国召开的新领军者年会，又称夏季达沃斯年会，与冬季达沃斯年会并蒂相映、相得益彰。当前中方与世界经济论坛合作势头良好，相信将来扩大合作的前景十分广阔。世界经济论坛将日益成为连接中国与世界的桥梁，成为增进世界对中国发展道路了解、认同与支持的重要平台。
抽取式 自动摘要	世界经济论坛将日益成为连接中国与世界的桥梁，成为增进世界对中国发展道路了解、认同与支持的重要平台每年1月，全球目光聚焦瑞士东部小镇达沃斯——来自世界各地的政、商、学界人士群贤毕至、济济一堂，共商天下大事。当今世界正经历深刻复杂变化，国际体系和国际秩序深度调整，国际力量对比深刻变化，各种新机遇新挑战层出不穷。值此之际，人们都想知道，中国对当前的“全球新局势”有何真知灼见，有何治理良方，中国会采取什么样的应对之道。人们还想知道，作为世界第二大经济体的中国，经济发展前景如何，在推动国际经济合作上将发挥什么样的作用。年会上的中国声音，必将增强国际社会对中国发展的信心，也会增强对世界经济前景的信心。中国通过全面深化改革，推动经济社会持续健康发展，将为世界提供更多机遇。中国推动建立以合作共赢为核心的新型国际关系，必将使全世界从中受益。
生成式 标准摘要	今年达沃斯年会的主题是“全球新局势”。年会上的中国声音，必将增强国际社会对中国发展的信心，也会增强对世界经济前景的信心。上世纪90年代初以来，中国的多位国家领导人曾出席达沃斯年会。每年9月世界经济论坛在中国召开的新领军者年会，又称夏季达沃斯年会，与冬季达沃斯年会并蒂相映、相得益彰。当下，达沃斯小镇银装素裹，国际媒体已开始将镜头聚焦小镇，世界对即将在达沃斯发出的中国声音充满期待。
生成式 自动摘要	世界经济论坛将日益成为连接中国与与世界的桥梁增进世界对中国的认可。

6 结语

自动摘要是将长文本提炼为简洁精炼的短文本的过程，能够帮助人们快速浏览文本资源并知晓文章大意，节省阅读成本的同时，也提高了知识利用效率，特别是在信息资源日益庞大的当下，自动摘要技术的需求更是与日俱增。笔者以 NEPD 中 2015 年 1 月、2015 年 6 月和 2016 年 1 月 3 个月的人民日报分词语料作为实验语料，根据新闻文本特征，面向人民日报语料设计了基于关键词词频排序和关键词簇排序的抽取式自动摘要算法，并构建了基于指针生成网络的生成式自动摘要模型，均在 Rouge 测评中取得了良好的实验结果，生成的摘要结果具有较好的完整性。笔者将在接下来的研究中完善算法，改进模型，增强模型的复用性，并对评价方法做出改进，加入文本内外部多个特征，增加人工生成标准摘要数据集和人工打分的环节，以提高自动摘要的流畅性和可读性。

参考文献：

[1] 王帅 , 赵翔 , 李博 , 等 .TP-AS: 一种面向长文本的两阶段自动摘要方法 [J]. 中文信息学报 , 2018, 32(6): 71-79.
[2] 黄水清 , 王东波 . 新时代人民日报分词语料库构建、性能及应用 (一)——语料库构建及测评 [J]. 图书情报工作 , 2019, 63(22): 5-12.
[3] 黄水清 , 王东波 . 基于人民日报语料的中央一号文件词频历时分析 [J]. 农业图书情报学报 , 2020, 32(3): 4-9.
[4] 莫燕 , 王永成 . 中文文献摘要的自动编制 [J]. 现代图书情报技术 , 1993(3): 10-12.
[5] 王永成 . 自动编制文献摘要及知识的自动提取 [J]. 现代图书情报技术 , 1993(3): 13-28.
[6] 王永成 , 许慧敏 .OA 中文文献自动摘要系统 [J]. 情报学报 , 1997(2): 49-53.
[7] 王知津 . 基于句子选择的自动文本摘要方法及其评价 [J]. 现代图书情报技术 , 1998(1): 46-51, 58.
[8] 史磊 , 王永成 . 英文文献自动摘要系统研究 [J]. 情报学报 , 1999(6): 504-508.
[9] 熊娇 , 王明文 , 李茂西 , 等 . 基于词项—句子—文档三层图模型的多文档自动摘要 [J]. 中文信息学报 , 2014, 28(6): 201-207.
[10] 张筱丹 , 胡学钢 . 基于向量空间模型的自动摘要冗余处理研究 [J]. 合肥工业大学学报 (自然科学版), 2010, 33(9): 1355-1358.
[11] 刘星含 , 霍华 . 基于互信息的文本自动摘要 [J]. 合肥工业大学学报 (自然科学版), 2014, 37(10): 1198-1203.
[12] 纪文倩 , 李舟军 , 巢文涵 , 等 . 一种基于 LexRank 算法的改进的自动文摘系统 [J]. 计算机科学 , 2010, 37(5): 151-154, 218.
[13] 曾哲军 . 基于连续 LexRank 的多文本自动摘要优化算法研究 [J]. 计算机应用与软件 , 2013, 30(10): 209-212, 245.
[14] 刘静 , 肖璐 . 基于依存句法分析的多主题文本摘要研究 [J]. 情报杂志 , 2014, 33(6): 167-171.
[15] 王帅 , 赵翔 , 李博 , 等 .TP-AS: 一种面向长文本的两阶段自动摘要方法 [J]. 中文信息学报 , 2018, 32(6): 71-79.
[16] 吴云 , 杨长春 , 梅佳俊 , 等 . 词句协同自动摘要提取方法 [J]. 计算机工程与设计 , 2018, 39(9): 2776-2779, 2810.
[17] 陈晨 , 张璐 , 伍之昂 . 词句协同排序的自动摘要算法 [J]. 江苏大学学报 (自然科学版), 2016, 37(04): 443-449.
[18] 丁建立 , 李洋 , 王家亮 . 基于双编码器的短文本自动摘要方法 [J]. 计算机应用 , 2019, 39(12): 3476-3481.
[19] 冯读娟 , 杨璐 , 严建峰 . 基于双编码器结构的文本自动摘要研究 [J]. 计算机工程 , 2020, 46(6): 60-64.
[20] 廖涛 , 刘宗田 , 王先传 . 基于事件的文本表示方法研究 [J]. 计算机科学 , 2012, 39(12): 188-191.
[21] 徐馨韬 , 柴小丽 , 谢彬 , 等 . 基于改进 TextRank 算法的中文文本摘要提取 [J]. 计算机工程 , 2019, 45(3): 273-277.
[22] 陈海华 , 黄永 , 张炯 , 等 . 基于引文上下文的学术文本自动摘要技术研究 [J]. 数字图书馆论坛 , 2016(8): 43-49.
[23] 黄水清 , 李志燕 , 梁刚 . 面向计算机类文献的自动摘要系统的研究与实现 [J]. 图书与情报 , 2006(3): 93-97.
[24] 张晗 , 赵玉虹 . 基于语义图的医学多文档摘要提取模型构建 [J]. 图书情报工作 , 2017, 61(8): 112-119.
[25] 陈志敏 , 姜艺 , 赵耀 . 基于用户查询扩展的自动摘要技术 [J]. 计算机应用研究 , 2011, 28(6): 2188-2190.
[26] 李芳 , 何婷婷 . 面向查询的多模式自动摘要研究 [J]. 中文信息学报 , 2011, 25(2): 9-14.
[27] 张哲铭 , 任淑霞 , 郭凯杰 . 结合主题感知与通信代理的文本摘要模型 [J]. 西安电子科技大学学报 , 2020, 47(3): 97-104.
[28] 陈燕敏 , 王晓龙 , 刘远超 , 等 . 一种基于文章主题和内容的自动摘要方法 [J]. 计算机工程与应用 , 2004(33): 11-14.
[29] 罗芳 , 汪竞航 , 何道森 , 等 . 融合主题特征的文本自动摘要方法研究 [J]. 计算机应用研究 , 2021, 38(1): 129-

133.
[30] 杜秀英 . 基于聚类与语义相似分析的多文本自动摘要方法 [J]. 情报杂志 , 2017, 36(6): 167-172.
[31] 吴世鑫 , 黄德根 , 李玖一 . 基于语义对齐的生成式自动摘要研究 [J]. 北京大学学报 (自然科学版), 2021, 57(1): 6.
[32] 方旭 , 过弋 , 王祺 , 等 . 核心词修正的 Seq2Seq 短文摘要 [J]. 计算机工程与设计 , 2018, 39(12): 3610-3615.
[33] 唐晓波 , 翟夏普 . 基于混合机器学习模型的多文档自动摘要 [J]. 情报理论与实践 , 2019, 42(2): 145-150.
[34] 谭金源 , 刁宇峰 , 祁瑞华 , 等 . 基于 BERT-PGN 模型的中文新闻文本自动摘要生成 [J]. 计算机应用 , 2021, 41(1): 127-132.
[35] 张克君 , 李伟男 , 钱榕 , 等 . 基于深度学习的文本自动摘要方案 [J]. 计算机应用 , 2019, 39(2): 311-315.
[36] 李维勇 , 柳斌 , 张伟 , 等 . 一种基于深度学习的中文生成式自动摘要方法 [J]. 广西师范大学学报 (自然科学版), 2020, 38(2): 51-63.
[37] 肖元君 , 吴国文 . 基于 Gensim 的摘要自动生成算法研究与实现 [J]. 计算机应用与软件 , 2019, 36(12): 131-136.
[38] 官礼和 .Internet 网络新闻文本自动摘要的研究 [J]. 计算机工程与设计 , 2007(14): 3518-3520, 3545.
[39] 韩永峰 , 许旭阳 , 李弼程 , 等 . 基于事件抽取的网络新闻多文档自动摘要 [J]. 中文信息学报 , 2012, 26(1): 58-66.
[40] 沈洲 , 王永成 , 许一震 , 等 . 一种面向新闻文献的自动摘要系统的研究与实践 [J]. 计算机工程 , 2000(9): 70-72.
[41] 李孟爽 , 昝红英 , 贾会贞 . 基于多特征和 Ranking SVM 的微博新闻自动摘要研究 [J]. 郑州大学学报 (理学版), 2017, 49(2): 44-48.
[42] 王凯祥 , 任明 . 基于查询的新闻多文档自动摘要技术研究 [J]. 中文信息学报 , 2019, 33(4): 93-100.
[43] 黄小江 , 万小军 , 肖建国 . 基于协同图排序的对比新闻自动摘要 [J]. 北京大学学报 (自然科学版), 2013, 49(1): 31-38.
[44] 柯修 , 王惠临 . 基于混合方法的多语言多文档自动摘要系统构建及实现 [J]. 图书馆学研究 , 2013(2): 66-72.
[45] 叶雷 , 余正涛 , 高盛祥 , 等 . 多特征融合的汉越双语新闻摘要方法 [J]. 中文信息学报 , 2018, 32(12): 84-91.
[46] 高永兵 , 王宇 , 马占飞 . 基于 CR-PageRank 算法的个人事件自动摘要研究 [J]. 计算机工程 , 2016, 42(11): 64-69.
[47] 陈卓群 , 王平 . 面向中文微博摘录式摘要方法研究 [J]. 情报科学 , 2015, 33(3): 130-134.
[48] 高永兵 , 钟振华 , 王宇 , 等 . 基于混合方法的中文微博自动摘要技术研究 [J]. 计算机工程与科学 , 2016, 38(6): 1257-1261.
[49] 贾晓婷 , 王名扬 , 曹宇 . 基于加权主题分布表达的微博文本摘要生成研究 [J]. 东北师大学报 (自然科学版), 2020, 52(1): 69-74.
[50] Text-Summarizer-Pytorh-Chinese[EB/OL]. [2021-07-07]. https://github.com/LowinLi/Text-Summarizer-Pytorch-Chinese.
[51] 彭敏 , 高斌龙 , 黄济民 , 等 . 基于高质量信息提取的微博自动摘要 [J]. 计算机工程 , 2015, 41(7): 36-42.
[52] LUHN H P. The automatic creation of literature abstracts[J]. IBM journal of research and development, 1958, 2(2): 159-165.
[53] 阮一峰 .TF-IDF 与余弦相似性的应用（三）：自动摘要 [EB/OL]. [2021-07-07]. http://www.ruanyifeng.com/blog/2013/03/automatic_summarization.html.
[54] SEE A, LIU P J, MANNING C D. Get to the point: Summarization with pointer-generator networks[J]. arXiv preprint arXiv:1704.04368, 2017.
[55] 程爽 . 浅谈《人民日报》改扩版的三个变化（3）[EB/OL]. [2021-07-07]. https://baike.baidu.com/redirect/7e44WWpuHPxVjlVjuIAMGFXvpzQ0nX6dtcm9N58nsqPgZqu9Xe51VC9kbRkCKxL7T3HLNLWACS5_clRah9xQ4caM3Wxuxf0d6PFTO7bT9zOcRDK1CYukrEXagCY.
[56] 俞士汶 , 朱学锋 , 段慧明 . 大规模现代汉语标注语料库的加工规范 [J]. 中文信息学报 , 2000(6): 58-64.
[57] 黄水清 , 王东波 . 国内语料库研究综述 [J]. 信息资源管理学报 , 2021, 11(3): 4-17, 87.
[58] 百度智能云 . 新闻摘要 [EB/OL]. [2021-07-07]. https://cloud.baidu.com/product/nlp_apply/news_summary.
[59] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Text summarization branches out, 2004: 74-81.

作者贡献说明：
梁　媛：进行数据处理，起草论文；
王东波：提供研究思路，设计研究方案；
黄水清：提出相关概念及整体研究思路，修订完稿。

知识管理论坛
2022 年 第 4 期（总第 40 期）
ChinaXiv合作期刊

Automatic Summary Generation of News for *People's Daily Online* Corpus

Liang Yuan[1,2] Wang Dongbo[1,2] Huang Shuiqing[1,2]
1. College of Information Management, Nanjing Agricultural University, Nanjing 210095
2. Research Center for Humanities and Social computing, Nanjing Agricultural University, Nanjing 210095

Abstract: [Purpose/significance] This paper conducts a study for the mainstream news media for *People's Daily Online* corpus, aiming to provide ideas and practical support for the study of automatic text summarization, which can then be applied to news and other related text information processing, and contribute to knowledge aggregation services and information access research. **[Method/process]** The experimental corpus of this research was the sub-corpus of the *People's Daily Online* in January 2015, June 2015 and January 2016 in the new era *People's Daily* (NEPD). Based on TF-IDF, Textrank and other extractive automatic summarization algorithms, based on the generative automatic abstractive summarization model for the pointer-generator network, the research was carried out and analyzed and evaluated the summarization results. **[Result/conclusion]** The experiment builds a news extraction automatic abstractive algorithm the Pointer-Generator Networks model for *the People's Daily* corpus, and constructs a network model of news generative automatic summary pointer generation for *People's Daily Online* corpus. Fruitful experimental results are evaluated by Rouge indicator (including 3 indicators: Rouge-1, Rouge-2 and Rouge-L). This article provides corpus support and practical support for the automatic news summarization system.

Keywords: *People's Daily* extractive automatic summarization generative automatic summarization NEPD pointer-generator networks